Fox Module 22 Conditional distribution of the response variable practice problems

(The attached PDF file has better formatting.)

** Exercise 22.1: Conditional distributions

A statistician chooses the conditional distribution of the response variable in a generalized linear model.

A. The choice of the conditional distribution may be intuitive (based on attributes of the response variable). What conditional distributions might be used for claim frequency, claim severity, and retention rates?
B. The choice of the conditional distribution may be empirical. How does the conditional distribution depend on the standard deviations of the values in each group?

*Part A:* A individual policyholder's claim frequency distribution is usually Poisson (or close to a Poisson). Each policyholder has an expected claim count, and the distribution of claims is Poisson. Claim frequency is claims divided by exposures, which are constant for the policyholder, so claim frequency is Poisson distributed.

The population claim frequency in heterogeneous classes is more skewed. Actuaries often use a negative binomial distribution, which is easy to work with. GLMs generally use Gamma distributions, which are of the exponential family.

*Jacob:* What does a heterogeneous class mean?

*Rachel:* If all policyholders in a class are identical, then the claim frequency distribution of the class is the claim frequency distribution of the policyholder. In practice, policyholders in a class differ.

*Illustration:* Most adolescent drivers have high motor insurance claim frequencies, but they differ greatly: some adolescents are careful drivers with low claim frequencies and others are aggressive drivers or poor drives with high claim frequencies. Each adolescent driver has a Poisson distribution for claim frequency. The class of adolescent drivers has a more skewed distribution for claim frequency.

*Jacob:* Do GLMs use Poisson distributions or negative binomial distributions?

*Rachel:* GLMs generally use over-dispersed Poisson distributions, which have a constant ratio of the variance to the mean (like the Poisson distribution), but this constant ratio is more than one.

*Intuition:* The critical item for fitting the GLM is the relation of the variance to the mean. For claim frequency, the variance at each point is roughly proportional to the expected value at that point.

*Jacob:* How do we determine the variance at a point?

*Rachel:* For discrete factors, we use the observed variance of the response variable in the class.

- *Claim frequency:* For claim frequency, the variance in each class is roughly proportional to the mean of the class. The examples below illustrate this property. For quantitative explanatory variables, we estimate the variance at each point with residual plots.

- *Claim severity:* For claim severity, the size-of-loss distribution is more skewed. Actuaries generally use Gamma, lognormal, or Pareto distributions. For claim severity, the variance at each point is roughly proportional to the square of the expected value at that point.

- *Retention rates:* Policyholder retention is a Bernoulli outcome: one renews or does not renew. Actuaries use binomial distributions: the variance is proportional to $\mu \times (1 - \mu)$, where $\mu$ is the expected value.

*Take heed:* GLMs use $\mu_j$ to denote the expected value of the conditional distribution at point *j*. If the response variable is a probability, we often use $\pi_j$. Retention rates, mortality rates, and response rates are probabilities.

*Part B:* We examine the means and standard deviations in each class. If the class dimensions are factors, such as age, sex, territory, and vehicle type (in motor insurance), policyholders are a multi-dimensional array. For this discussion, each cell in the array is a class.

*Illustration:* Youthful vs mature are levels of the factor *age*. Male vs female are levels of the factor *sex*. Urban vs rural are levels of the factor *territory*. Youthful rural females are a class, as are mature, urban, males. We examine the means and variances of the members of each class.

For simplicity, suppose the class system has two factors with two levels for each factor: sex (male vs female) and territory (urban vs rural). The insured population has four classes:

- Class #1: male-urban
- Class #2: male-rural
- Class #3: female-urban
- Class #4: female-rural

The conditional distribution for the response variable depends on the ratio of the standard deviation to the mean. We estimate the means and standard deviations for each class. Below are illustrations of three cases.

*Illustration:* A 2 × 2 rating system has four classes, with means and standard deviations as shown below.

| *Class:* | *Class 1* | *Class 2* | *Class 3* | *Class 4* |
|---|---|---|---|---|
| *Mean:* | 800 | 500 | 400 | 200 |
| *Standard Deviation:* | 185 | 115 | 95 | 45 |
| *Ratio:* | 0.23125 | 0.23 | 0.2375 | 0.225 |

The ratio of the standard deviation to the mean is steady, so the variance is proportional to the square of the mean. We use a Gamma distribution for the GLM.

*Jacob:* Do we use the ratio of the variance to the *average observed value* or to the *fitted mean* in the GLM?

*Rachel:* We want to use the ratio of the variance to the fitted mean in the GLM.

- When we start fitting the GLM, we don't know the fitted values.
  - The average observed value is a proxy for the fitted mean.
- After we fit the GLM, we re-inspect the ratio of the variance to the fitted mean to test if the conditional distribution is reasonable.

*Jacob:* Can you give an example when one would choose a Poisson distribution?

*Rachel:* A 2 × 2 rating system has four classes, with means and standard deviations as shown below.

| *Class:* | *Class 1* | *Class 2* | *Class 3* | *Class 4* |
|---|---|---|---|---|
| *Mean:* | 16% | 10% | 8% | 4% |
| *Standard Deviation:* | 40% | 32% | 28% | 20% |
| *Variance* | 0.16 | 0.1024 | 0.0784 | 0.0400 |
| *Ratio:* | 1.000 | 1.024 | 0.980 | 1.000 |

The ratio of the variance to the mean is constant. One would choose a Poisson distribution for the GLM.

*Jacob:* In this illustration, the variance equals the mean, as is true for a Poisson distribution. What if the variance is proportional to the mean but not equal to it?

*Rachel:* We still use a Poisson distribution. If the classes are homogeneous, the variances may be equal to the means. If the classes are heterogeneous, the variances will generally be greater than the means.

*Jacob:* Can you give an example when one would choose a binomial distribution?

*Rachel:* A 2 × 2 rating system has four classes, with means and standard deviations as shown below.

| *Class:* | *Class 1* | *Class 2* | *Class 3* | *Class 4* |
|---|---|---|---|---|
| *Mean:* | 50% | 70% | 85% | 95% |
| *Standard Deviation:* | 50% | 46% | 36% | 22% |
| *Variance* | 0.25 | 0.2116 | 0.1296 | 0.0484 |
| *Mean × (1 – Mean)* | 0.25 | 0.21 | 0.1275 | 0.0475 |
| *Ratio:* | 1.0000 | 1.0076 | 1.0165 | 1.0189 |

The ratio of the variance to (the mean × the complement of the mean) is constant. One would choose a Poisson distribution for the GLM.

** Exercise 22.2: Conditional distribution

The range of the dependent variable depends on the conditional distribution in the generalized linear model.

What are the ranges of the following conditional distributions? Is the distribution symmetric or skewed? If the distribution is symmetric, is it heavy-tailed, light-tailed, or neither?

A. Gaussian (normal) distribution
B. Binomial distribution
C. Poisson distribution
D. Gamma distribution
E. Lognormal distribution

*Part A:* The Gaussian (normal) distribution has a range of $(-\infty, +\infty)$. The distribution is symmetric, not skewed. The terms heavy-tailed and light-tailed are in relation to the normal distribution, which (by definition) is neither.

*Jacob:* Is the normal distribution symmetric about zero?

*Rachel:* It is symmetric about its mean.

*Jacob:* We use normal distributions for many random variables that do not take negative values or which have skewed distributions.

*Rachel:* By the central limit theorem, if the response variable is the sum of many similarly sized distributions of any sort, it has an asymptotic normal distribution.

*Intuition:* If the response variable is the sum of N independent random variables of similar size, even if they are not identically distributed, its distribution approaches a normal distribution as N approaches infinity.

Actuaries compute average values for claim frequency. claim severity, and retention rates for large samples.

- The distribution for one policyholder may be Poisson, Gamma, or binomial.
- The distribution of the average can often be approximated by a normal distribution.

Generalized linear models with appropriate conditional distributions are better than regression analysis with normal distributions for most data sets, but the approximations with regression analysis are often good.

*Part B:* The binomial distribution has a range of $( (0, 1, \ldots, n_j) / n_j)$, where $n_j$ is the number of exposures.

*Illustration:* Retention rates are the number of policyholders who renew divided by total policyholders . With $n_j$ policyholders, 0 to $n_j$ may renew. (General insurance uses *renewal rate* instead of *retention rate*.) Studies of new medication ask whether the patient recovers from illness or does not recover (or dies vs does not die).

Fox refers to some response variables as hidden probabilities. Percentages, test scores with multiple choice questions, bond defaults; and mortality rates are all probabilities.

*Illustration:* A multiple choice exam has fifty questions with five options each. The distribution of scores for candidates with absolutely no knowledge is a binomial distribution with N = 50 and $\pi$ = 20%.

*Part C:* The Poisson distribution has a range of (0, 1, 2, …). Claim counts may be 0, 1, 2, … (any integer).

*Illustration:* Claim frequencies are claim counts divided by exposures, sot hey are also Poisson distributions.

*Jacob:* How does general insurance claim frequency differ from life insurance mortality rates?

*Rachel:* A life insurance policyholder can die only once, so mortality rates have Bernoulli distributions. Drivers may have several accidents the same year. If the occurrence of one accident is independent of the occurrence of other accidents, claim frequency has a Poisson distribution.

*Parts D and E:* The Gamma distribution and the lognormal distribution have ranges of $(0, +\infty)$. Most important for GLMs, they are positively skewed. Above some minimal value, the likelihood of a value X is greater than the likelihood of a value Z if $Z > X$.

For some Gamma distributions, this minimal value is zero. The likelihood is monodically decreasing over the entire range (the positive real numbers). For other Gamma distributions and all lognormal distributions, the mode of the distribution is more than zero.

The lognormal distribution is for stock prices and for insurance claim severities. Pareto distributions are also used for claim severities.

*Jacob:* Fox doesn't discuss the lognormal or Pareto distributions. This seems strange: the Black-Scholes formula assumes stock prices have a lognormal distribution, and reinsurance actuaries commonly use Pareto distributions for loss cost distributions.

*Rachel:* The lognormal and Pareto distributions are not in the exponential family of distributions. The Gamma distribution is similar to these distribution and is in the exponential family. Generalized linear models use Gamma distributions instead of lognormal or Pareto distributions.

*Jacob:* The Gamma, lognormal, and Pareto distributions have different tails. Modern computers can do maximum likelihood estimates using many distributions. Why not use the distribution that fits the data best?

*Rachel:* The important item is the relative variance as a function of the expected value. Gamma. lognormal, and Pareto distributions have variances proportional to the square of the mean. Maximum likelihood estimation gives similar result for these distributions.

See Fox, *Regression analysis*, Chapter 15, Structure of Generalized linear models, page 381